Journal of Biomedical Informatics 117 (2021) 103763

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data

Jaime Lynn Speiser [*]

Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

ARTICLE INFO

Keywords:
Random forest
Feature selection
Variable selection
Longitudinal outcomes
Clustered outcomes
Binary mixed model forest

ABSTRACT

Background: Machine learning methodologies are gaining popularity for developing medical prediction models for datasets with a large number of predictors, particularly in the setting of clustered and longitudinal data. Binary Mixed Model (BiMM) forest is a promising machine learning algorithm which may be applied to develop prediction models for clustered and longitudinal binary outcomes. Although machine learning methods for clustered and longitudinal methods such as BiMM forest exist, feature selection has not been analyzed via data simulations. Feature selection improves the practicality and ease of use of prediction models for clinicians by reducing the burden of data collection. Thus, feature selection procedures are not only beneficial, but are often necessary for development of medical prediction models. In this study, we aim to assess feature selection within the BiMM forest setting for modeling clustered and longitudinal binary outcomes.

Methods: We conducted a simulation study to compare BiMM forest with feature selection (backward elimination or stepwise selection) to standard generalized linear mixed model feature selection methods (shrinkage and backward elimination). We also evaluated feature selection methods to develop models predicting mobility disability in older adults using the Health, Aging and Body Composition Study dataset as an example utilization of the proposed methodology.

Results: BiMM forest with backward elimination generally offered higher computational efficiency, similar or higher predictive performance (accuracy and area under the receiver operating curve), and similar or higher ability to identify correct features compared to linear methods for the different simulated scenarios. For predicting mobility disability in older adults, methods generally performed similarly in terms of accuracy, area under the receiver operating curve, and specificity; however, BiMM forest with backward elimination had the highest sensitivity.

Conclusions: This study is novel because it is the first investigation of feature selection for developing random forest prediction models for clustered and longitudinal binary outcomes. Results from the simulation study reveal that BiMM forest with backward elimination has the highest accuracy (performance and identification of correct features) and lowest computation time compared to other feature selection methods in some scenarios and similar performance in other scenarios. Many informatics datasets have clustered and longitudinal outcomes and results from this study suggest that BiMM forest with backward elimination may be beneficial for developing medical prediction models.

1. Introduction

Developing prediction models for longitudinal outcomes is a common goal in many medical studies. Advances in computer science and technology allow researchers to collect a multitude of data over time, such as electronic health record data and large cohort study data. For example, the Health, Aging and Body Composition Study (Health ABC) aimed to assess longitudinal measures of mobility disability in a cohort

Abbreviations: Health ABC, Health, Aging, and Body Composition; GLMM, generalized linear mixed model; BiMM, binary mixed model; AUC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; DGP, Data generating process; MCC, Matthews correlation coefficient; ROC, receiver operating characteristic.

[*] Address: Wake Forest School of Medicine, Department of Biostatistics and Data Science, Division of Public Health Sciences, Medical Center Boulevard, Winston-Salem, NC 27157, USA.

E-mail address: jspeiser@wakehealth.edu.

https://doi.org/10.1016/j.jbi.2021.103763
Received 15 October 2020; Received in revised form 3 March 2021; Accepted 23 March 2021
Available online 26 March 2021

of older adults [1]. Longitudinal data arises when outcomes and/or predictors are collected repeatedly over time throughout a study, and this structure may create correlation within a subject because observations for the same subject are dependent. The more general term for correlated data is clustered data, in which observations are correlated among a group. An example of this is data collected for several family members because these observations are likely correlated. Modeling data with a clustered or longitudinal structure needs to accommodate the potential for dependence among correlated observations. Methods used to develop prediction models should account for longitudinal and clustered outcomes to accurately reflect correlation within a dataset. While standard modeling techniques provide a basis for model development involving a moderate numbers of predictors, advances in methodology are needed to address the challenges of modern longitudinal datasets with many potential predictor variables.

Machine learning methodologies are gaining popularity for developing medical prediction models for datasets with a large number of predictors [2–15], particularly in the setting of clustered and longitudinal data. Although clustered and longitudinal data arise from different scenarios, methods to accommodate these data are similar. Several methods have been proposed for longitudinal and clustered outcomes, including support vector machine [16], neural networks [17], decision tree [18–31], and random forest [32–34] frameworks. The main idea behind these methods is to incorporate machine learning models into the generalized linear mixed model (GLMM) framework to account for correlation among clusters and longitudinal outcomes [16–34]. Many methods employ a process similar to the Expectation-Maximization algorithm in which a machine learning model is fit, followed by a GLMM fitted with results or output from the machine learning model, and then the outcome is updated based on the fitted GLMM [16–34]. This process is iteratively repeated until convergence is reached. We focus on binary outcomes in this study because many medical studies use binary (two-class) endpoints which dynamically change between states over time. We previously developed a method to implement random forest for clustered and longitudinal outcomes called Binary Mixed Model (BiMM) forest [34].

Although machine learning methods for clustered and longitudinal methods exist, none consider feature selection. Reducing the number of features, or predictor variables, within a model is often an essential part of developing clinical decision support tools since researchers want to eliminate superfluous features which are inconsequential for predicting outcomes. This is especially important when there is a large number of potential predictors (e.g., cohort data such as Health ABC or electronic health record data) or when there is a significant cost associated with the collection of certain predictors (e.g., tests, questionnaires, or laboratory measurements which are not collected in a standard clinic visit). For example, developing a longitudinal prediction model for mobility disability in older adults, there is a significant increase in efficiency of prediction when the number of features is decreased, even by just a few measures [35]. Feature selection improves the practicality and ease of use of models for clinicians by reducing the burden of data collection. For these reasons, feature selection procedures are not only beneficial, but are often necessary for development of prediction models in real world applications.

Despite recent popularity of machine learning methods for clustered and longitudinal outcomes, feature selection for these methods remains understudied. A preprint paper focused on random forests for high dimensional continuous longitudinal outcomes suggests using stepwise feature selection, but provides no simulation study or justification for this recommendation [36]. A recent study by Calhoun and colleagues [37] incorporate feature selection into a repeated measures random forest framework; however, this study has limitations because the performance of the method was not established in a simulation study and the performance for the data application is moderate (area under the receiver operating curve of 0.622). In this paper, our objective is to determine the best feature selection method for BiMM forest using a simulation study and real-world data application and compare performance, accuracy of features identified and computation time to existing methods (namely, GLMM with coefficient shrinkage and GLMM with backward elimination). A novel aspect of the current paper is that we present the improvement of the performance of BiMM forest including a feature selection method for modeling both clustered and longitudinal data.

In this paper, we discuss feature selection within the BiMM forest setting for modeling clustered and longitudinal binary outcomes. The framework is presented in the methods section of the paper (Section 2). We conduct a simulation study to compare predictive performance, accuracy of model specification, and computation time for both standard and proposed feature selection for BiMM forest models (Section 3). Finally, we present an application of the feature selection methods using Health ABC data to develop a prediction model for risk of mobility disability in older adults over time (Section 3).

2. Methods

2.1. Binary mixed model (BiMM) forest

The main idea of BiMM forest is to incorporate random forest into the GLMM framework [38]. For binary outcomes, GLMMs have the form

$$\text{logit}(y_{it}) = \boldsymbol{X_{it}\beta} + Z_{it}b_{it},$$

where y_{it} is the binary outcome for cluster $i = 1,\ldots,M$ for longitudinal measurements $t = 1,\ldots,T_i$, logit() is the logistic link function, $\boldsymbol{X_{it}}$ is a matrix of fixed covariates for cluster i for longitudinal measurement t, $\boldsymbol{\beta}$ is a vector of fitted coefficients for the fixed covariates, Z_{it} is the clustered covariate for cluster i for longitudinal measurement t, and b_{it} is the fitted random effect for cluster i for longitudinal measurement t. In words, the predictors are partitioned into fixed and random effects which are linearly related to the binary outcome through the logistic link function. Because GLMMs require assumptions which are not always valid (e.g. linear association between the predictors and the outcome through the link function and interactions must be specified by the user), we developed a method called BiMM forest [34].

The algorithm is depicted in Fig. 1 in the dotted oval portion of the diagram. To summarize, BiMM forest uses an algorithm in which a random forest is developed (Step 1), then the predicted probability of each observation from the random forest is used within a Bayesian GLMM to adjust for longitudinal outcomes (Step 2). This portion of the algorithm takes the form

$$\text{logit}(y_{it}) = \beta_0 + \beta_1 \text{RF}(\boldsymbol{X_{it}}) + Z_{it}b_{it},$$

where RF($\boldsymbol{X_{it}}$)is represented within the GLMM as the predicted probability from the random forest of each longitudinal observation $t = 1,\ldots,T_i$ for cluster $i = 1,\ldots,M$. β_0 is the coefficient for the intercept an β_1 is the coefficient for the vector of random forest probabilities, RF($\boldsymbol{X_{it}}$). The updating step of the BiMM forest involves employing a split function that takes the predicted probabilities from the GLMM model and makes them binary (0/1) to update outcomes (Step 3). BiMM forest can be compiled as a single-iteration method (where the algorithm stops after the first run) or as an algorithmic method where the binary outcome is updated at each iteration (repeat Steps 1–3 until convergence). In this study, we focus of the split function proposed in the original BiMM forest paper called H3, which updates the outcome without favoring model sensitivity or specificity [34].

Within a simulation study [34] and a data application for developing a longitudinal model for the condition of acute liver failure patients [39], we previously demonstrated that BiMM forest offers comparable or slightly higher prediction accuracy compared to other models.

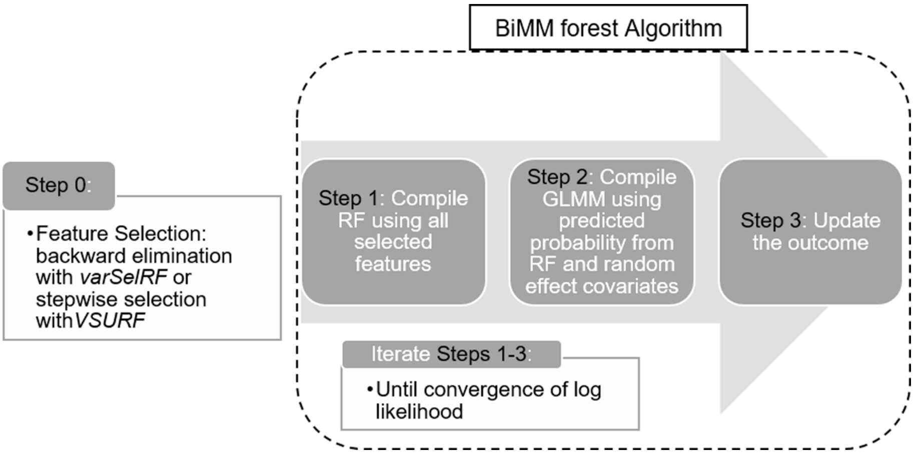
BiMM forest Algorithm
Step 0:
•Feature Selection: backward elimination with varSelRF or stepwise selection withVSURF
Step 1: Compile RF using all selected features
Step 2: Compile GLMM using predicted probability from RF and random effect covariates
Step 3: Update the outcome
Iterate Steps 1-3:
•Until convergence of log likelihood

Fig. 1. The BiMM forest algorithm with feature selection as a pre-step before development of a prediction model.

2.2. BiMM forest with feature selection

The goal of the present study is to provide guidance on feature selection when using the BiMM forest framework for developing prediction models. Several methods for feature selection with random forest exist [11,40–49]. A recent study demonstrated that methods implemented in the R packages *varSelRF* [42] and *VSURF* [43] may be preferable for binary outcomes due to parsimony, accuracy and computational efficiency [50]. Both of these approaches are wrapper methods. Hence, we investigate using these methods prior to modeling in order to determine which features should be included in BiMM forest.

varSelRF [42] is a method based on backward elimination of features proposed by Diaz-Uriarte et al. which removes the least important features while maintaining a similar error rate to the full model (containing all features). The method begins by developing a random forest model and determining variable importance measures (the user can specify to use mean decrease Gini or the default is permutation importance). Next, features are removed from the model according to the original ordering of variable importance and the out-of-bag error rate is computed. Features are removed successively until the out-of-bag error rate is significantly different from the rate from the full model containing all predictors. This was one of the first feature selection methods for random forest proposed in the literature and remains a widely used method for random forest feature selection, particularly for large datasets such as microarray data.

VSURF [43] is a method based on a stepwise procedure proposed by Genuer et al. which also maintains a similar error rate to the full model. The method begins by developing many random forests (e.g., 50) and averaging the permutation variable importance across the random forest models to create an ordering of most important to least important features. The least important features are removed from consideration after this first step, based on a threshold value of variable importance. Next, features are added into the random forest model one at a time, only if the error decrease is larger than a threshold. The features of the last model are then selected.

We propose a pre-step to BiMM forest modeling in which feature selection is done prior to developing the prediction model for the clustered or longitudinal binary outcome (Fig. 1). In Step 0, features to be included in the model are determined by either backward elimination via *varSelRF* [42] or stepwise selection via *VSURF* [43] and then included in the BiMM forest method (Steps 1–3). Feature selection is done as a pre-step outside of the BiMM forest method in order to maximize computational efficiency and simplicity of the method. Because feature selection is done as a pre-step to BiMM forest and random effects are incorporated within the BiMM forest method, the longitudinal nature of the features is not explicitly considered during selection.

2.3. Design of simulation study

We conduct a simulation study based on data in which the relationship between predictors and outcome is known in order to assess the performance of the feature selection with BiMM forest. The simulation setup is the same as the original BiMM forest paper [34], based on a real world longitudinal dataset of 1064 patients which represent the clusters. We simulate cluster sizes (repeated measurements) of 2, 4 and 7. In total, there are six continuous features and five binary features. Data are split at the cluster-level, meaning that if a cluster (patient) is selected for the training data then all observations for that cluster (patient) are included in the training data. Similarly, if a cluster (patient) is selected for the testing data then all observations for that cluster (patient) are included in the testing data. From the 1064 clusters available in the simulated data, we randomly select 100 clusters for model development (training data) and 500 clusters for model validation (testing data) which are independent of those used to develop the model for each simulation run. Thus, our testing data consists of independent clusters of observations from the training data. We chose to use 100 clusters, representing subjects/participants, for developing the models because this is a challenging modeling scenario and therefore is likely to highlight differences between methods. More clusters (500) were used for model validation to obtain predictions for a variety of different observations to evaluate how the models performed. We conduct 100 simulation runs.

To determine outcomes within the simulation, we use three data generating processes for the fixed portion of the outcome: a linear structure, a tree structure, and a complex structure. The linear data generating process includes two continuous features and one binary feature as predictors which are linearly related to the binary outcome using the logistic link function. The tree data generating process includes the same three features as the linear process, but the outcome is calculated using a series of binary splits on the features. The complex data generating process uses five decision trees with four continuous features and three binary features to determine outcome based on majority voting of the trees. However, one of these features is intentionally omitted from model development and feature selection to add complexity to the data generation process. After the fixed portion of the outcome is determined, it is added onto random effects which account for correlation among the clusters. We use a small and large random effect, generated from standard normal distributions centered at zero with standard deviations of 0.1 and 0.5 respectively. The outcome for each observation is derived by adding the fixed (from the linear, tree or complex data generating process) and random portions and using a cut point to create a binary endpoint. More details about outcome derivation are included in the original BiMM forest paper [34].

2.4. Methods to be compared

We compare several existing methods to the newly proposed feature selection methods for BiMM forest with the simulated data. For a benchmark, we develop models with no feature selection using standard GLMMs with the R package *lme4* [51], BiMM forest with one iteration and BiMM forest with H3 updates. We conduct feature selection using GLMM LASSO with the R package *glmmLasso* [52] and GLMM with backward elimination with the R package *StatisticalModels* [53]. These are compared to BiMM forest (1 iteration and H3 updated models) with feature selection done as a pre-step using backward elimination with *varSelRF* [42] and stepwise selection with *VSURF* [43]. All models are implemented using R software with default parameter values. We used R version 3.6.1 on a computer with an Intel Core i7-7700 CPU 3.6 GHz, 3600 Mhz, 4 Cores, 8 Logical processors and 16.0 GB of RAM.

2.5. Evaluation metrics

For each simulation run and model, we calculate several evaluation metrics. The primary evaluation metric is area under the receiver operating curve (AUC) for all models to assess overall performance using the R package *ROCR* [54]. Prediction accuracy (accuracy of predictions for the testing dataset), F1 score and Matthews correlation coefficient (MCC) are additionally used to evaluate performance of the models and are presented in Appendix A. Since in the simulation study the association between predictors and outcome is known, we are able to assess the accuracy of features selected by the methods. To evaluate the accuracy of feature selection, we record the number of features selected in total, the number of features correctly selected, and the percent of correct models (defined as the percent of simulation runs in which the exact features used to derive the outcome for the data generating process are selected). Finally, we calculate the computation time for the models. We

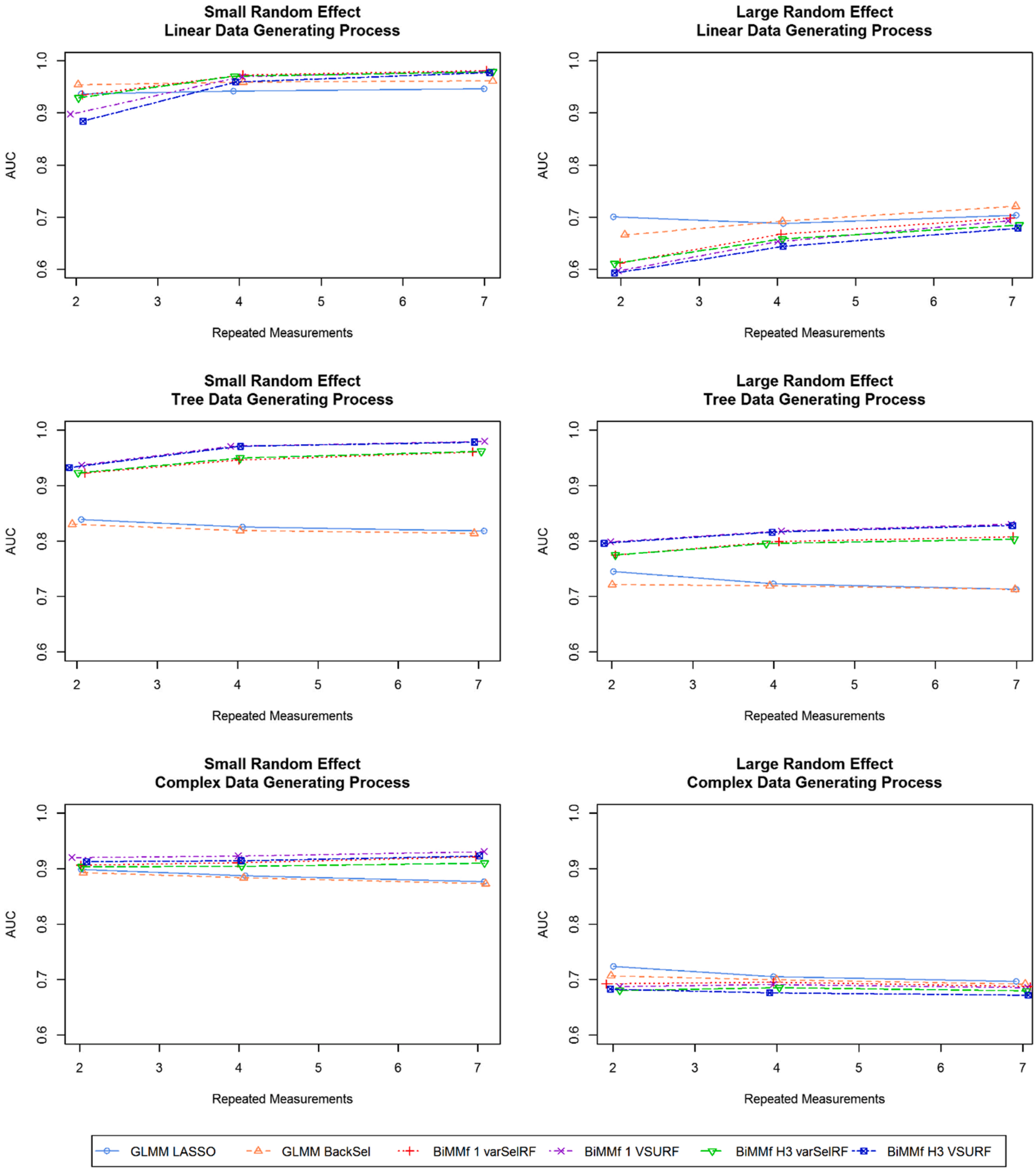

Fig. 2. Mean AUC of methods for the simulated scenarios across 100 runs. **Legend:** These plots display AUC across the number of repeated measurements for each of the simulated scenarios. Higher values of AUC represent better model performance.

compare the models using t tests for evaluation metrics averaged across all of the simulated data generating processes.

3. Results

3.1. Model performance

Model performance was assessed using AUC (Fig. 2, Table A.1) and prediction accuracy (Fig. A.1, Table A.1). In general, across the data generating processes, cluster sizes (repeated measures) and random effect sizes, accuracy and AUC of methods including feature selection was similar to that of methods using all features. Of the methods including features selection, most had similar prediction accuracy and AUC for the linear data generating process and the complex data generating process, aside from the linear data generating process with a large random effect and a cluster size of 2 repeated measurements. In this scenario, the GLMM LASSO and GLMM with backward selection had higher AUC compared to BiMM forest methods with feature selection. However, the BiMM forest methods with feature selection had higher accuracy and AUC compared to linear models with feature selection (GLMM LASSO and GLMM with backward selection) for the tree data generating process. Of the BiMM forest feature selection methods for the tree data generating process, models using *VSURF* for feature selection had higher accuracy and AUC compared to models using *varSelRF* for feature selection, particularly for the small random effect scenarios. BiMM forest models with H3 updates had similar accuracy and AUC compared to BiMM forest models using one iteration for the same feature selection method (e.g., comparing BiMM forest with 1 iteration for *varSelRF* and

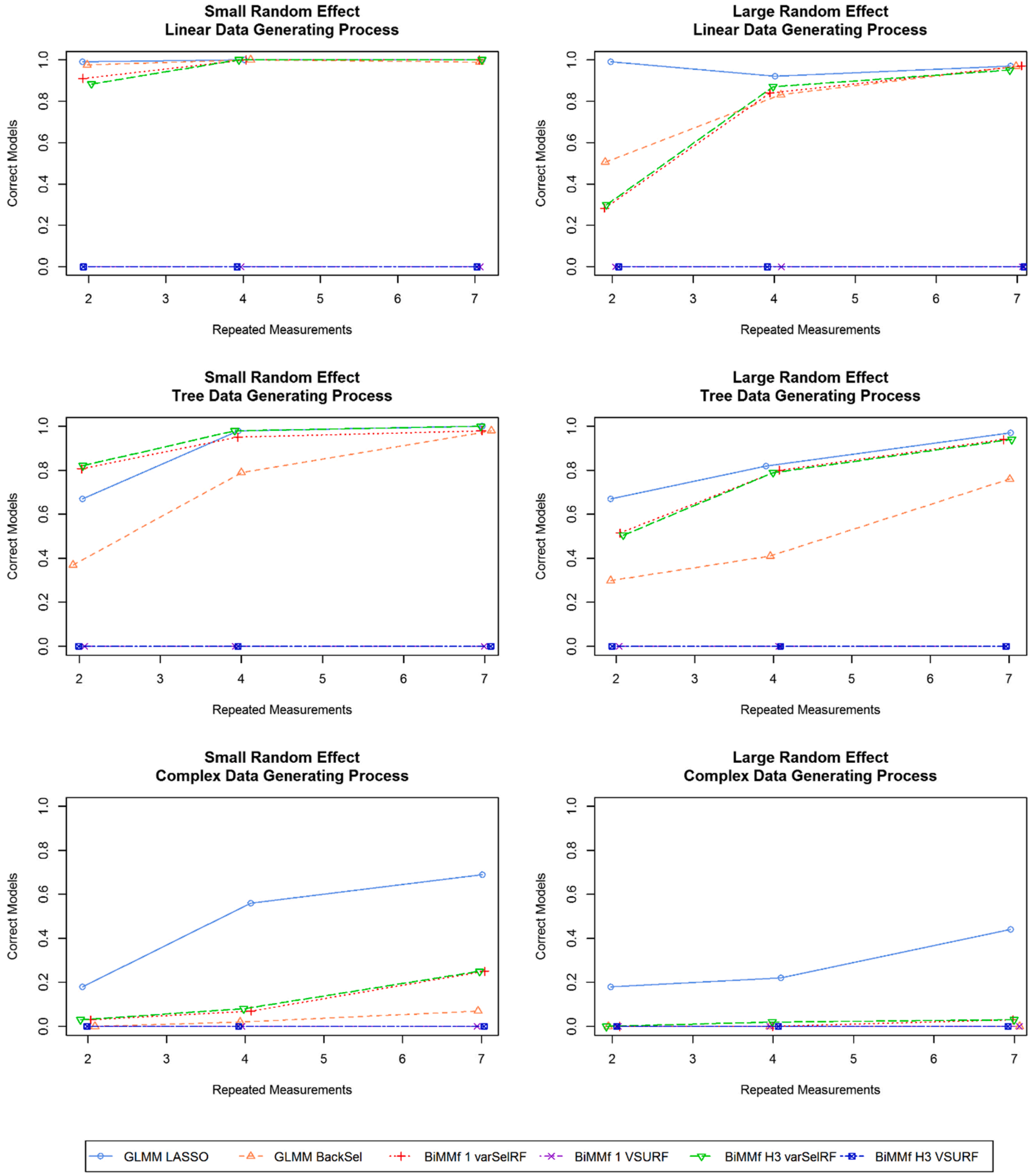

Fig. 3. Percent of correctly specified models of methods for the simulated scenarios across 100 runs. **Legend:** These plots display the percent of correctly specified models (all features correctly identified) across the number of repeated measurements for each of the simulated scenarios. Higher values of the percent of correctly specified models represent better model performance.

VSURF; comparing BiMM forest with H3 updates for *varSelRF* and *VSURF*). Performance of models developed without feature selection (i.e., standard GLMM, BiMM forest with 1 iteration, and BiMM forest with H3 updates without) were similar to their counterparts with feature selection, suggesting that reducing the features did not impact predictive ability (Table A.1, Table A.2 and Table A.4).

In addition to accuracy and AUC, models were assessed using the F1 score (Fig. A.2, Table A.2) and MCC (Fig. A.3, Table A.2). Performance of the models using F1 score is similar to that of AUC and accuracy described above: generally, for the linear and complex data generating processes the F1 scores were similar, whereas for the tree data generating process BiMM forest methods had better performance compared to GLMM LASSO and GLMM with backward selection. These results were also similar for the MCC statistic.

3.2. Accuracy of features selected

In addition to assessing performance statistics, we assessed the feature selection methods in terms of their ability to accurately select the correct features across the 100 simulation runs (Fig. 3, Table A.3). In general, the methods tended to select more features than were included in the data generating processes (Table A.3, total number of feature selected). For the linear and tree data generating processes, 3 features were used. In the complex data generating process, 7 features were used to generate outcomes but one of these was intentionally omitted from modeling and feature selection, therefore we define a correctly specified model as containing the 6 features included in the feature selection phase. The BiMM forest methods with *VSURF* feature selection fared the worst, with no models being correctly specified, although the method did identify some of the features correctly. Aside from these, the

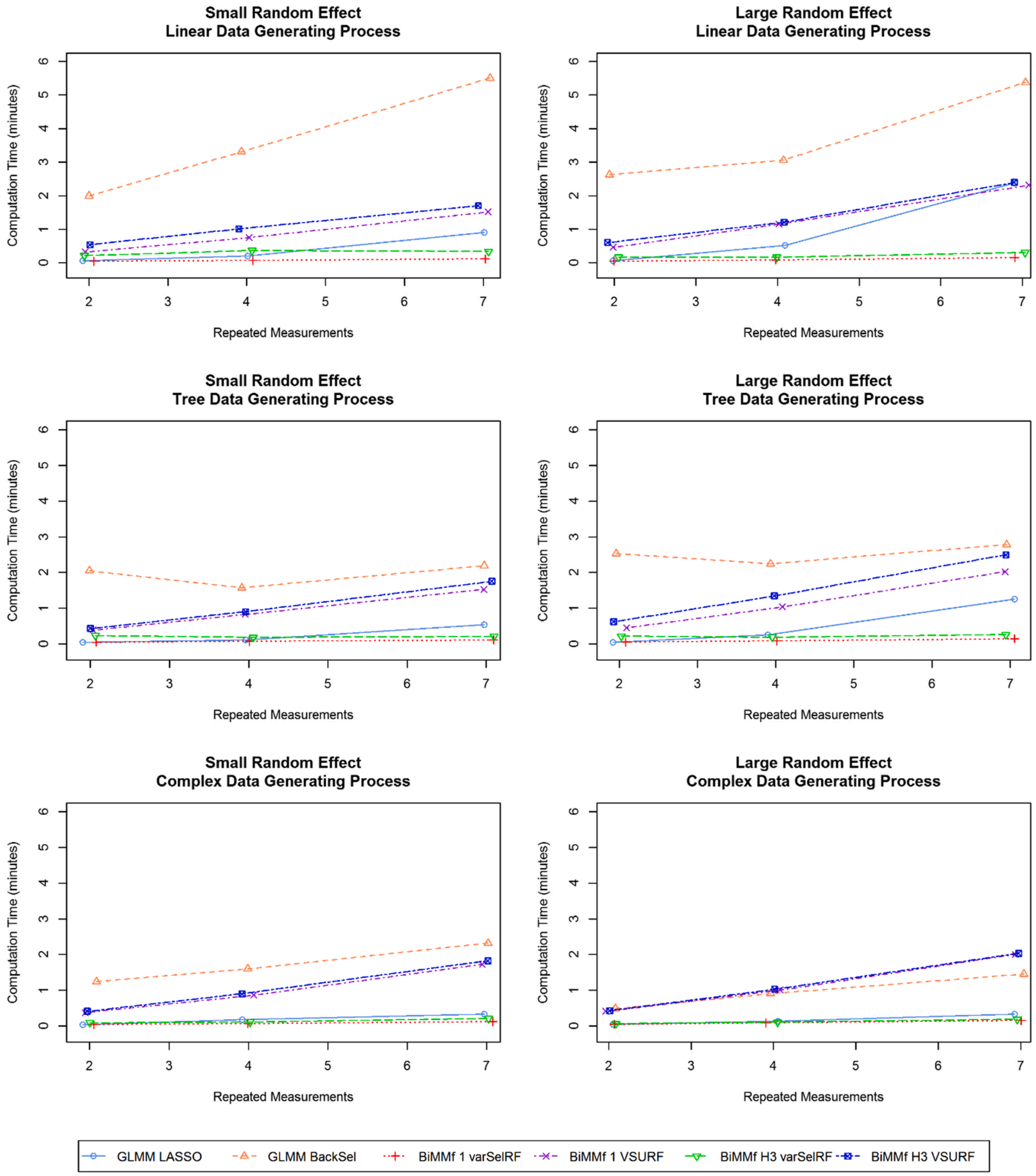

Fig. 4. Mean computation time (minutes) of methods for the simulated scenarios across 100 runs. **Legend:** These plots display computation time in minutes across the number of repeated measurements for each of the simulated scenarios. Lower values of computation time represent preferable models.

remaining methods were able to correctly select features for the linear data generating process with a small random effect. For the linear data generating process with a large random effect, the GLMM LASSO was able to correctly identify features better than other methods for a cluster size of 2, but for cluster sizes of 4 and 7, the methods performed similarly. For the tree data generating processes, BiMM forest methods with *varSelRF* and GLMM LASSO were able to identify features correctly a higher proportion of times compared to other methods across the simulation runs. For the complex data generating process, all methods struggled to select the correct features, though the GLMM LASSO had the best percent of models correctly specified (ranging from 20 to 60% for different simulated scenarios).

3.3. Computation time

Computation time (minutes) was assessed for the methods within the data simulation (Fig. 4, Table A.4). Across all simulated scenarios, GLMM with backward selection had the highest run time. BiMM forest methods (1 iteration and H3 updates) with *VSURF* feature selection had moderate computation time compared to other methods. The most efficient (least amount of time) methods were GLMM LASSO and BiMM forest with *varSelRF* for feature selection using either 1 iteration and H3 updates.

3.4. Data application

To illustrate the use of BiMM forest with feature selection and compare it to standard methods for a real application, we develop prediction models for identifying older adults at-risk of mobility disability using longitudinal measures from the Health ABC dataset [1]. This study began in 1997 and followed 3075 well-functioning, community-dwelling men and women aged 70–79 at baseline. Requests for a copy of the dataset may be made through the National Institute on Aging webpage (https://healthabc.nia.nih.gov/). The primary endpoint for our study is mobility disability (yes/no), defined as self-reported inability to walk 1/4 mile or climb up 10 steps. Mobility disability was collected during the study follow-up at years 2 to 6, 8 and 10 from enrollment. The predicted probabilities from the models were used to determine correct and incorrect predictions of outcome: if the predicted probability was greater than 0.5 then the prediction was mobility disability and if the predicted probability was less than 0.5 then the prediction was non-mobility disabled. We define sensitivity as the proportion of correct model predictions for older adults reporting mobility disability and specificity as the proportion of correct model predictions for older adults not reporting mobility disability.

We consider both fixed and longitudinal predictors of mobility disability for development of the prediction models, based on previous studies of risk factors [55–59]. Fixed predictors collected at baseline include race, gender, age, marital status, number of people living in the household, education, income, and lifestyle variables (e.g., smoking, drinking, sleeping data). Longitudinal predictors collected at years 2 to 6, 8, and 10 include age at the visit time, body mass index, comorbidities in the past year (hypertension, osteoarthritis, cardiovascular disease, cancer, diabetes and stroke), physical performance measures (Epese score for standing balance, chair stands, and ease walking), and muscle strength self-reported data (ease lifting/carrying 10 lb, ease standing from chair without using arms). In total, we preform feature selection using 20 fixed predictors and 16 longitudinal predictors to develop the models for mobility disability. While a total of 36 features is not a huge number of predictors for modeling, any reduction in features needed to be included in the model would result in significant increases in efficiency of prediction because of the reduced time and effort to collect data. This is particularly the case for features which require additional tests or measures that are non-standard in clinic visits (e.g., physical performance measures). All features considered for modeling were previously identified as meaningful for predicting mobility disability in older adults, and the problem here was to reduce the number of features to be included in a prediction model.

Data was collected for predictors and outcome at years 2–6, 8 and 10, or until the participant died. There were on average 8 years of data for the participants in the study. In total, 3.5% (671/19029) of observations had missing outcome data (i.e., mobility disability is missing at a certain visit) and are therefore excluded from analysis. For missing predictor data, we use the R package *missForest* to impute missing values, which does not consider the longitudinal nature of the data structure [60]. The amount of missing data was minimal for most predictors (less than 15% for most predictor variables, see Appendix B). Inspection of missing values by outcome for baseline and longitudinal (repeated measures) data revealed that for some features, there was more missing data in the mobility disability group compared to the non-mobility disability group (Appendix B HABC dataset description file). After the missing data were imputed, we implement a random split of the entire dataset into a training (N = 2050 participants, with M = 12288 observations total) and test (N = 1025 participants, with M = 6070 observations total) dataset to assess the predictive accuracy of the model for new subjects.

A complete description of the Health ABC dataset may be found elsewhere in the literature [1]. For this study, baseline and longitudinal data are described in Appendix B. The cohort consisted of 1235 (42%) black older adults and 1551 (52%) female older adults. Mean age at baseline was 74 (standard deviation 3). The outcome variable, mobility disability, was imbalanced throughout the follow up. Overall, the event rate for mobility disability was 4.2% (762/18358). Nothing specifically was done to address the imbalanced outcome for developing prediction models in our study.

We compare the following methods for the Health ABC data application: GLMM LASSO, GLMM with backward selection, and BiMM forest (1 iteration and H3 updates) with feature selection using *varSelRF* and *VSURF*. The GLMM LASSO selected a total of 23 features, whereas the GLMM with backward selection selected 12. BiMM forest models with *varSelRF* used 11 features and models with *VSURF* used 6 features. These are a significant reduction compared to the 36 features entered into the model.

Performance statistics and 95% binomial confidence intervals for the methods are presented in Table 1 for the test dataset. Exact binomial confidence intervals are calculated using the base R function *binom.test()* and AUC and its confidence intervals are calculated using the R package *cvAUC* [61]. We include results for a standard random forest model with all features, which ignores the clustered, longitudinal structure of the data. The standard random forest had similar AUC, accuracy and specificity compared to other models which use feature selection, but slightly lower sensitivity compared to BiMM forest models with *varSelRF* feature selection. BiMM forest models with *varSelRF* feature selection had the highest AUC and accuracy, although all models had high AUC (ranged from 0.95 to 0.98) and accuracy (ranged from 96 to 99%). All models had similar specificity, which was close to 1, whereas performance varied for sensitivity. The BiMM forest model with 1 iteration and *varSelRF* feature selection had the highest sensitivity (82%) compared to other models. Receiver operating characteristic (ROC) curves and precision-recall curves are presented in Fig. 5. Consistent with AUC results, most models had similar ROC curves, aside from the GLMM LASSO which was slightly worse. The BiMM forest models had the best precision recall curves compared to the GLMM models, with BiMM forest *varSelRF* models performing slightly better than those with *VSURF* feature selection.

4. Discussion

In this paper, we analyze feature selection as a pre-step to the BiMM forest method for modeling clustered and longitudinal binary outcomes. A novel contribution of our study is that it is the first analysis of feature selection for developing random forest models for clustered and longitudinal binary outcomes. Our data simulation investigated using

Table 1
Results for the data application.

Method	# Features	AUC	Accuracy	Sensitivity	Specificity
Standard Random Forest	36	0.984 (0.972,0.995)	0.989 (0.986,0.992)	0.731 (0.669,0.788)	0.999 (0.998,1)
GLMM LASSO	23	0.950 (0.943,0.956)	0.955 (0.951,0.959)	0.120 (0.0934,0.150)	0.993 (0.991,0.994)
GLMM Backward Selection	12	0.980 (0.974,0.986)	0.980 (0.977,0.982)	0.546 (0.503,0.589)	0.999 (0.999,1)
BiMM forest 1 Iteration varSelRF	11	0.983 (0.967,0.998)	0.992 (0.990,0.994)	0.815 (0.758,0.863)	0.999 (0.998,1)
BiMM forest 1 Iteration VSURF	6	0.98 (0.965,0.996)	0.986 (0.983,0.989)	0.771 (0.711,0.824)	0.994 (0.992,0.996)
BiMM forest H3 Updates varSelRF	11	0.973 (0.947,0.998)	0.990 (0.987,0.992)	0.753 (0.692,0.808)	0.999 (0.998,1)
BiMM forest H3 Updates VSURF	6	0.958 (0.925,0.99)	0.984 (0.98,0.987)	0.639 (0.573,0.701)	0.997 (0.995,0.998)

Notes: For AUC, accuracy, sensitivity and specificity, the proportions are presented along with 95% confidence intervals. # features: the number of features that the method selected for inclusion in the model.

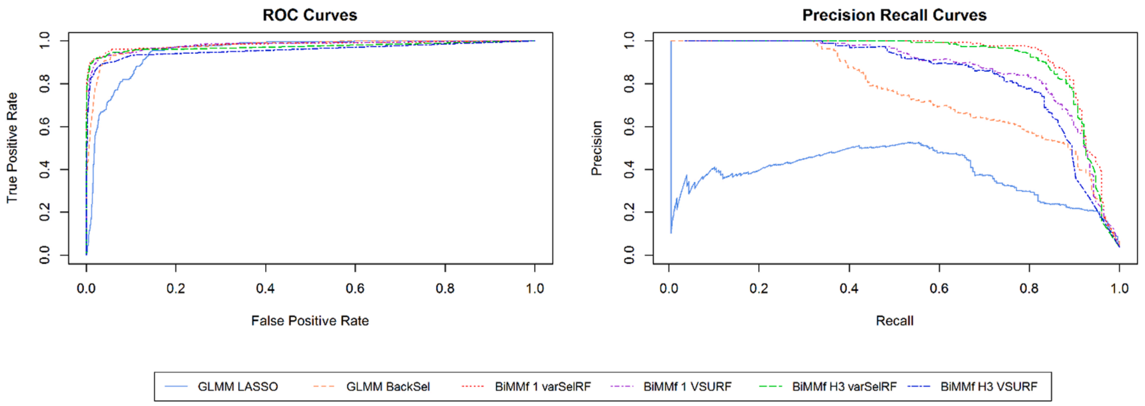

Fig. 5. ROC curves and precision recall curves for the Health ABC dataset.

varSelRF and *VSURF* for feature selection prior to modeling with BiMM forest and compared this to GLMM LASSO and GLMM with backward selection. Our simulation study suggests that BiMM forest (either 1 iteration or with H3 updates) with *varSelRF* feature selection may be the best model for developing a prediction model for clustered and longitudinal binary outcomes because it optimized highest accuracy (performance and identification of correct features) and lowest computation time. For the linear and tree data generating processes, the BiMM forest methods with *varSelRF* had similar proportions of correctly specified models compared to GLMM LASSO. None of the methods were able to identify the correct models for the complex data generating process particularly well, although GLMM LASSO had higher proportions of correctly specified models compared to the other methods. A novel aspect of this study is the inclusion of computation time as an evaluation metric. GLMM with backward selection had the highest run times, which is not surprising because removing features one at a time requires building many mixed models and much computational expense. BiMM forest models with *varSelRF* had the lowest computation times, consistent with previous studies that have shown the speed of the method [50]. Computation time increased linearly with cluster size for BiMM forest models with *VSURF* and for GLMM with backward selection, whereas computation time remained similar regardless of cluster size for BiMM forest models with *varSelRF* and GLMM LASSO.

For the data application to develop a longitudinal prediction model for mobility disability risk over time using Health ABC data, the method with the highest accuracy and AUC was BiMM forest with 1 iteration and *varSelRF* feature selection. This method also had much higher sensitivity (82%) compared to GLMM models (12% for the GLMM LASSO and 55% for the GLMM with backward selection), while all models had high specificity (greater than 99%). This is important because it means that the BiMM forest with 1 iteration and *varSelRF* feature selection can best identify individuals that may be at risk for developing mobility disability compared to the other methods. However, sensitivity and specificity values could be altered by using different threshold values based on the ROC plots. Aside from potential increases in performance, another motivation for using BiMM forest over standard random forest is that it can accommodate dependencies induced by longitudinal or clustered data, whereas standard random forest does not explicitly handle longitudinal or clustered data.

Combining simulation study and data application results in terms of performance (AUC and accuracy), ability to select correct features and models, and computation time, we recommend using BiMM forest (either 1 iteration or H3 updates) with *varSelRF* feature selection for development of parsimonious prediction models with clustered and longitudinal binary outcomes. Despite having an imbalanced outcome, BiMM forest models generally performed well. For datasets with imbalanced outcome when standard models do not provide good performance, a weighted sampling scheme may be adopted in which the rarer outcome class may be oversampled within the random forest portion of the BiMM forest method. This is a common way of addressing imbalanced outcomes. Models proposed are useful for prediction outcome in older adults for new subjects (independent from those included in model development). Future studies should evaluate the clinical utility of the models (i.e., if something like this were integrated into the electronic health record system, how useful would it be for clinicians and how would clinicians use these predictions to help make decisions regarding care for their patients).

Recently, a few approaches for random forests for clustered and longitudinal outcomes have been proposed [32–34], but few have investigated feature selection for prediction modeling. There are some methods for feature selection for clustered and longitudinal outcomes [62–69]; however, many of these are designed for specific dataset structures or types (e.g., continuous outcomes and/or features only; or all binary features) and are not appropriate for all datasets. Additionally, many of these methods require assumptions which are not always valid, such as linearity between features and outcome through the link function. Another limitation of these methods is that they require specification of interaction terms among features, which may not be known *a priori*. A final limitation of these methods is that many of them are computationally inefficient compared to machine learning approaches, such as BiMM forest. The BiMM forest approach with *varSelRF* feature selection overcomes many of the limitations of other methods because it can be used for datasets with diverse structures, naturally models nonlinear associations of features and outcome, can handle complex interactions among features without user specification, and has better computational efficiency compared to other methods.

Despite these benefits of BiMM forest with *varSelRF* feature selection, there are some limitations of our study which should be discussed. Firstly, our stimulation study focused on a synthetic dataset with just over 1000 clusters (participants) with cluster sizes of 2, 4 and 7, consequently these results may not generalize to other datasets with a larger number of clusters or with larger cluster sizes. A small number of features (11 total) were used in this simulation study to compare performance of methods in a setting where associations between features and outcome are specified. Most real-world applications will have more features; however, we used this design for the simulation study because of computational expense: larger datasets take longer to develop prediction models, and because many clinical datasets are around this size. A future study could investigate feature selection for clustered and longitudinal outcomes with big data or high dimensional datasets. A limitation of the data application is that there were some missing values of features, and these were imputed using standard random forest methodology which ignores clustering. In the future, we plan to analyze the impact of missing data on development of BiMM forest prediction models. We imputed all missing data before splitting into training and testing data, which may have inflated performance statistics for all models. Additionally, missingness varied by the outcome variable, suggesting a missing data pattern that may be non-ignorable. However, this is a difficult problem to address and since missing data imputation was not the focus of our study, we used standard methodology that assumes a missing at random mechanism. A future study should investigate the impact of the missingness mechanism on imputation for longitudinal prediction modeling. In a real-world application of these models, the last repeated measurement observation is most important because the previous observations are considered historical data. It is possible that the performance of predictions for the last repeated measure may be worse compared to the predictions for the other repeated measures. This was observed in the original BiMM forest paper data application [34], so it is plausible this occurred in the models here as well. The differences in performance should be considered when using these models in practice.

Finally, we propose feature selection as a pre-step to modeling with BiMM forest; however, an alternative method may be developed which incorporates feature selection within the BiMM forest method that would consider the clustered structure of the data, a strategy implemented on one previous method for continuous outcomes [36]. Performing feature selection without considering random effects may potentially remove important variables with high inter-cluster variability. Although this may provide better performance or accuracy of model specification, we decided not to do this in our study because conducting feature selection prior to modeling with BiMM forest still resulted in models which had good overall performance, and we wanted to keep the method as simple as possible. Using feature selection and BiMM forest modeling as independent steps may have resulted in overestimation of performance in the simulation study; however, results from the data application are consistent with the stimulation study in terms of performance, so the overestimation of performance in the simulation study is likely minor. Future studies should evaluate the generalizability of BiMM forest with feature selection for additional datasets and different scenarios. For example, it would be interesting to analyze performance of methods in the presence of cluster dependent features.

5. Conclusion

The main objective of this study was to analyze feature selection prior to modeling with BiMM forest for development of parsimonious prediction models for clustered and longitudinal binary outcomes. Our simulation study and data application suggest that features selected using *varSelRF* used in BiMM forest may be preferable because of high AUC and accuracy, good ability to identify correct features for inclusion in a model, and low computation time compared to competing methods. Many informatics datasets have clustered and longitudinal outcomes and results from this study suggest that BiMM forest with backward elimination may be beneficial for developing medical prediction models as an alternative to standard regression methods (e.g., GLMM).

An R package for conducting feature selection for use within BiMM forest is being developed and will be available on GitHub. R code implementing feature selection and BiMM forest methodology are available within a supplemental file (Appendix C).

6. Ethics approval and consent to participate

This study was exempt by the Institutional Review Board at Wake Forest School of Medicine.

7. Availability of data and materials

The dataset for the simulation study are described in a paper by Speiser et al [34] and are available by request from the Acute Liver Failure Study Group (repository.nikkd.nih.gov/studies/aalf/). The dataset for the data application in this study are available by request from the National Institute on Aging (healthabc.nia.nih.gov). R code implementing BiMM forest with feature selection is included within Appendix C.

Author contributions

JLS is the sole author of this study. She conceived the idea, developed the method, wrote and edited the manuscript.

CRediT authorship contribution statement

Jaime Lynn Speiser: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Institutes of Health National Center for Advancing Translational Sciences Grant (KL2 TR001421). The sponsor of this work had no involvement in the study design, analysis or writing of the report. This research was supported by National Institute on Aging (NIA) Contracts N01-AG-6-2101; N01-AG-6-

2103; N01-AG-6-2106; NIA grant R01-AG028050, and NINR grant R01-NR012459. This research was funded in part by the Intramural Research Program of the NIH, National Institute on Aging.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103763.

References

[1] L. Fredman, J.A. Cauley, S. Satterfield, E. Simonsick, S.M. Spencer, H.N. Ayonayon, et al., Caregiving, mortality, and mobility decline: The health, aging, and body composition (Health ABC) study, Arch. Intern. Med. 168 (2008) 2154–2162.

[2] A.L. Boulesteix, S. Janitza, J. Kruppa, et al., Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, Wiley Interdisciplinary Rev.: Data Min. Knowledge Discovery 2 (2012) 493–507.

[3] L. Breiman, Random forests, Mach Learn. 45 (2001) 5–32.

[4] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. Perez-Sanchez, J. A. Benediktsson, A. Thapa, et al., Automatic selection of molecular descriptors using random forest: Application to drug discovery, Expert Syst. Appl. 72 (2017) 151–159.

[5] D.R. Cutler, T.C.J. Edwards, K.H. Beard, et al., Random forest for classification in ecology, Ecology 88 (2007) 2783–2792.

[6] B.A. Goldstein, E.C. Polley, F. Briggs, Random forests for genetic association studies, Statist. Appl. Genet. Mol. Biol. 10 (2011) 1–34.

[7] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, D. Shen, Initiative AsDN. Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, Neurobiol. Aging 46 (2016) 180–191.

[8] B. Larivière, D. Van den Poel, Predicting customer retention and profitability by using random forests and regression forests techniques, Expert Syst. Appl. 29 (2005) 472–484.

[9] D.S. Siroky, Navigating random forests and related advances in algorithmic modeling, Statist. Surveys 3 (2009) 147–163.

[10] J.L. Speiser, V.L. Durkalski, W.M. Lee, Random forest classification of etiologies for an orphan disease, Stat. Med. 34 (2015) 887–899.

[11] V. Svetnik, A. Liaw, C. Tong, T. Wang, Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules, Multiple Classifier Systems: Springer, 2004, p. 334–343.

[12] R. Tang, J.P. Sinnwell, J. Li, D.N. Rider, M. de Andrade, J.M. Biernacka, Identification of Genes and Haplotypes that Predict Rheumatoid Arthritis using Random Forests, BMC proceedings: BioMed Central Ltd, 2009, p. S68.

[13] W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, et al., Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings Bioinf. (2012) bbs034.

[14] J.J. van der Zande, M. Dauwan, E. Van Dellen, P. Scheltens, A.W. Lemstra, C. J. Stam, Applying random forest machine learning to diagnose Alzheimer's disease and dementia with Lewy bodies: A combination of electroencephalography (EEG), clinical parameters and biomarkers, Alzheimer's & Dementia: J. Alzheimer's Assoc. 12 (2016) P661–P662.

[15] Q. Zhou, W. Hong, L. Luo, F. Yang, Gene selection using random forest and proximity differences criterion on DNA microarray data, J. Convergence Inform. Technol. 5 (2010) 161–170.

[16] J. Luts, G. Molenberghs, G. Verbeke, S. Van Huffel, J.A. Suykens, A mixed effects least squares support vector machine model for classification of longitudinal data, Comput. Stat. Data Anal. 56 (2012) 611–628.

[17] Y. Xiong, H.J. Kim, V. Singh, Mixed effects neural networks (menets) with applications to gaze estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, p. 7743–7752.

[18] M. Abdolell, M. LeBlanc, D. Stephens, R. Harrison, Binary partitioning for continuous longitudinal data: categorizing a prognostic variable, Stat. Med. 21 (2002) 3395–3409.

[19] A. Ciampi, Q. Lin, G. Yousif, GLIMTREE: RECPAM Trees with the Generalized Linear Model, Compstat: Springer, 1990, pp. 21–26.

[20] G. De'Ath, Multivariate regression trees: a new technique for modeling species-environment relationships, Ecology 83 (2002) 1105–1117.

[21] A. Dine, D. Larocque, F. Bellavance, Multivariate trees for mixed outcomes, Comput. Stat. Data Anal. 53 (2009) 3795–3804.

[22] S.-H. Eo, H. Cho, Tree-structured mixed-effects regression modeling for longitudinal data, J. Computat. Graph. Statist. (2013).

[23] M. Fokkema, N. Smits, A. Zeileis, T. Hothorn, H. Kelderman, Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees, Behavior Res. Methods 1–19 (2015).

[24] W. Fu, J.S. Simonoff, Unbiased regression trees for longitudinal and clustered data, Comput. Stat. Data Anal. 88 (2015) 53–74.

[25] A. Hajjem, F. Bellavance, D. Larocque, Mixed effects regression trees for clustered data, Statist. Probability Lett. 81 (2011) 451–459.

[26] A. Hajjem, D. Larocque, F. Bellavance, Generalized mixed effects regression trees, Statist. Probability Lett. 126 (2017) 114–118.

[27] Lee S. Keon, On generalized multivariate decision tree by using GEE, Comput. Stat. Data Anal. 49 (2005) 1105–1119.

[28] M.R. Segal, Tree-structured methods for longitudinal data, J. Am. Stat. Assoc. 87 (1992) 407–418.

[29] R.J. Sela, J.S. Simonoff, RE-EM trees: a data mining approach for longitudinal and clustered data, Mach Learn. 86 (2012) 169–207.

[30] J.L. Speiser, B.J. Wolf, D. Chung, C.J. Karvellas, D.G. Koch, V. Durkalski, BiMM tree: A decision tree method for modeling clustered and longitudinal binary outcomes, Commun. Statist. - Simulation Comput. (2018) 1–20.

[31] H. Zhang, Y. Ye, A tree-based method for modeling a multivariate ordinal response, Statist. Interface 1 (2008) 169.

[32] A. Hajjem, F. Bellavance, D. Larocque, Mixed-effects random forest for clustered data, J. Stat. Comput. Simul. 84 (2014) 1313–1328.

[33] C. Ngufor, H. Van Houten, B.S. Caffo, N.D. Shah, R.G. McCoy, Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c, J. Biomed. Inform. 89 (2019) 56–67.

[34] J.L. Speiser, B.J. Wolf, D. Chung, C.J. Karvellas, D.G. Koch, V.L. Durkalski, BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes, Chemometrics Intell. Lab. Syst. (2019).

[35] J.L. Speiser, K.E. Callahan, D.K. Houston, J. Fanning, T.M. Gill, J.M. Guralnik, et al., Machine learning in aging: an example of developing prediction models for serious fall injury in older adults, J. Gerontol.: Series A. 2020.

[36] L. Capitaine, R. Genuer, R. Thiébaut, Random forests for high-dimensional longitudinal data, arXiv preprint arXiv:190111279. 2019.

[37] P. Calhoun, R.A. Levine, J. Fan, Repeated measures random forests (RMRF): Identifying factors associated with nocturnal hypoglycemia, Biometrics (2020).

[38] P. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, Analysis of longitudinal data, Oxford University Press, 2002.

[39] J.L. Speiser, C.J. Karvellas, B.J. Wolf, D. Chung, D.G. Koch, V.L. Durkalski, Predicting daily outcomes in acetaminophen-induced acute liver failure patients with machine learning techniques, Comput. Methods Programs Biomed. 175 (2019) 111–120.

[40] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347.

[41] H. Deng, G. Runger, Gene selection with guided regularized random forest, Pattern Recogn. 46 (2013) 3483–3489.

[42] R. Díaz-Uriarte, S.A. De Andres, Gene selection and classification of microarray data using random forest, BMC Bioinf. 7 (2006) 3.

[43] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, VSURF: an R package for variable selection using random forests, The R Journal. 7 (2015) 19–33.

[44] A. Hapfelmeier, K. Ulm, A new variable selection approach using random forests, Comput. Stat. Data Anal. 60 (2013) 50–69.

[45] H. Ishwaran, U. Kogalur, Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.6. URL http://CRAN R-projectorg/package=randomForestSRC. 2014.

[46] S. Janitza, E. Celik, A.-L. Boulesteix, A computationally fast variable importance test for random forests for high-dimensional data, Adv. Data Anal. Classif. 1–31 (2015).

[47] H. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, et al., Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes, BMC Bioinf. 5 (2004) 81.

[48] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.

[49] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (2010) 1–13.

[50] J.L. Speiser, M.E. Miller, J. Tooze, E. Ip, A comparison of random forest variable selection methods for classification prediction modeling, Expert Syst. Appl. (2019).

[51] D. Bates, M. Maechler, B. Bolker, S. Walker, R.H.B. Christensen, H. Singmann, et al. Package 'lme4'. 2015.

[52] A. Groll, glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation. R package version 2017; 1: 25.

[53] T. Newbold, R package: StatisticalModels. Functions for generating, analysing, checking and plotting statistical models, 2020.

[54] T. Sing, O. Sander, N. Beernwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics 21 (2005).

[55] T.M. Gill, C.S. Williams, M.E. Tinetti, Assessing risk for the onset of functional dependence among older adults: the role of physical performance, J. Am. Geriatr. Soc. 43 (1995) 603–609.

[56] J.M. Guralnik, L. Ferrucci, E.M. Simonsick, M.E. Salive, R.B. Wallace, Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability, N. Engl. J. Med. 332 (1995) 556–562.

[57] J.M. Guralnik, L.P. Fried, M.E. Salive, Disability as a public health outcome in the aging population, Annu. Rev. Public Health 17 (1996) 25–46.

[58] T. Manini, Development of physical disability in older adults, Curr. Aging Sci. 4 (2011) 184–191.

[59] M. Pahor, J.M. Guralnik, W.T. Ambrosius, S. Blair, D.E. Bonds, T.S. Church, et al., Effect of structured physical activity on prevention of major mobility disability in older adults: the LIFE study randomized clinical trial, JAMA 311 (2014) 2387–2396.

[60] D.J. Stekhoven, missForest: Nonparametric Missing Value Imputation using Random Forest. 2013. R package version. 2019; 1.

[61] E. LeDell, M. Petersen, M. van der Laan, M.E. LeDell, Package 'cvAUC', 2014.

[62] H.D. Bondell, A. Krishna, S.K. Ghosh, Joint variable selection for fixed and random effects in linear mixed-effects models, Biometrics. 66 (2010) 1069–1077.

[63] A. Groll, G. Tutz, Variable selection for generalized linear mixed models by L 1-penalized estimation, Stat. Comput. 24 (2014) 137–154.

[64] J.G. Ibrahim, H. Zhu, R.I. Garcia, R. Guo, Fixed and random effects selection in mixed effects models, Biometrics. 67 (2011) 495–503.

[65] X. Ni, D. Zhang, H.H. Zhang, Variable selection for semiparametric mixed models in longitudinal studies, Biometrics. 66 (2010) 79–88.

[66] J. Schelldorfer, P. Bühlmann, S.V. De Geer, Estimation for high-dimensional linear mixed-effects models using ℓ1-penalization, Scand. J. Stat. 38 (2011) 197–214.

[67] J. Schelldorfer, L. Meier, P. Bühlmann, Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using ℓ1-penalization, J. Computat. Graph. Stat.. 23 (2014) 460–477.

[68] Y. Tang, H.J. Wang, Z. Zhu, Variable selection in quantile varying coefficient models with longitudinal data, Comput. Stat. Data Anal. 57 (2013) 435–449.

[69] P. Zhao, L. Xue, Variable selection in semiparametric regression analysis for longitudinal data, Ann. Inst. Stat. Math. 64 (2012) 213–231.